Optimized Framework based on Rough Set Theory for Big Data Preprocessing in Certain and Imprecise Contexts

Chelly Dagdia, Zaineb

Publication date:
2018

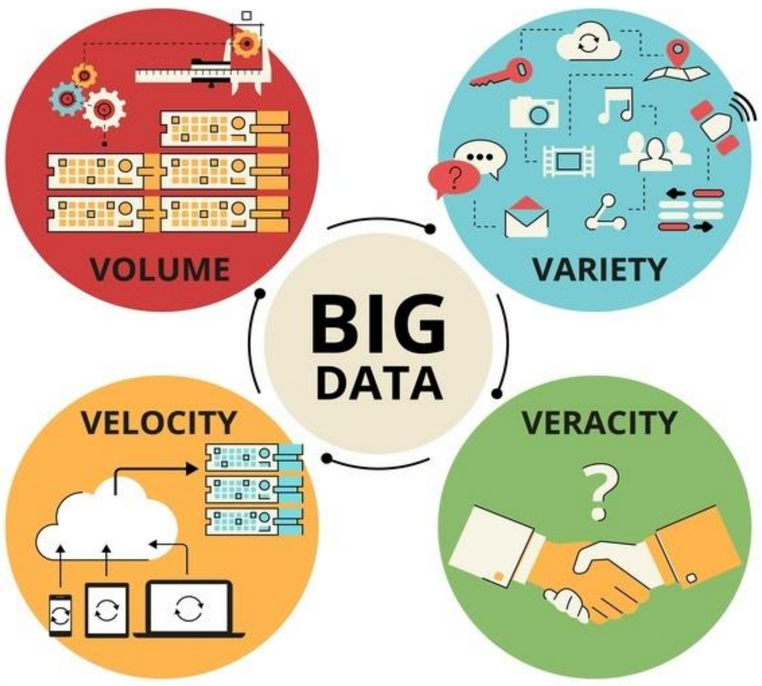
Citation for published version (APA):
Chelly Dagdia, Z. (2018). *Optimized Framework based on Rough Set Theory for Big Data Preprocessing in Certain and Imprecise Contexts*. Poster session presented at The 5th MCAA Annual Conference and General Assembly, Leuven, Belgium.

2018 MCAA Annual Conference and General Assembly
2nd-3rd February 2018, Leuven, Belgium

Optimized Framework based on Rough Set Theory for Big Data Preprocessing in Certain and Imprecise Contexts

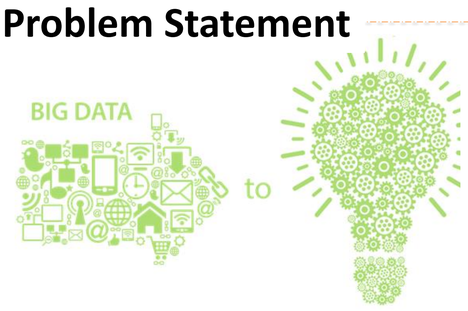
Zaineb Chelly Dagdia
Department of Computer Science, Aberystwyth University, Aberystwyth, United Kingdom

Context



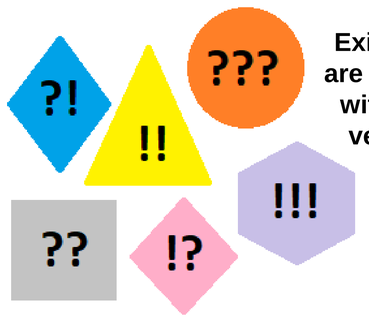
Motivation and Problem Statement

It has become difficult to quickly acquire the most useful information from the huge amount of data at hand.



Existing methods for big data preprocessing require additional information about the given data for thresholding and noise levels to be specified

Existing methods involve experts/users for parameterization

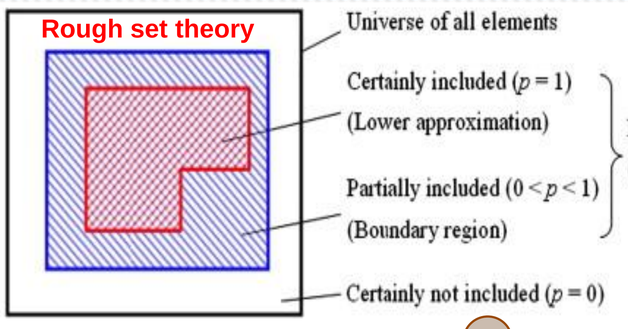


Existing methods are not able to deal with the big data veracity aspect

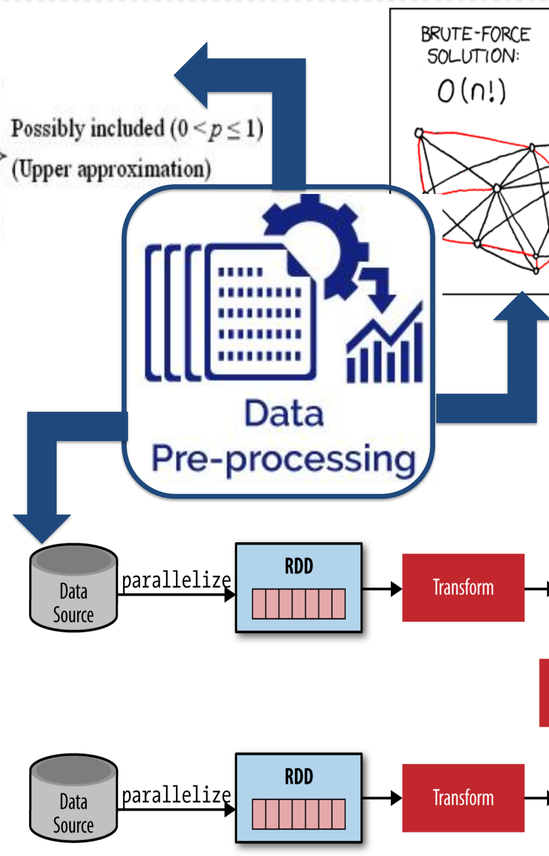
Existing methods are not able to deal with the big data computational requirements

Imperfection in data (imprecision, uncertainty, inconsistency).

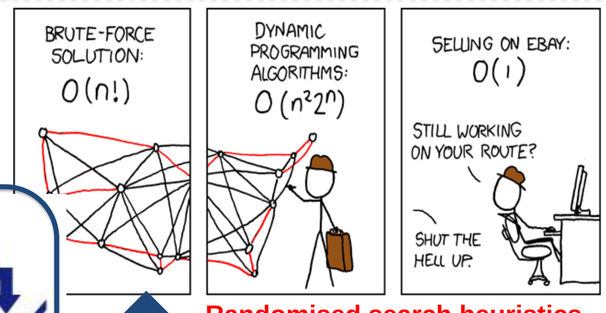
Project Description



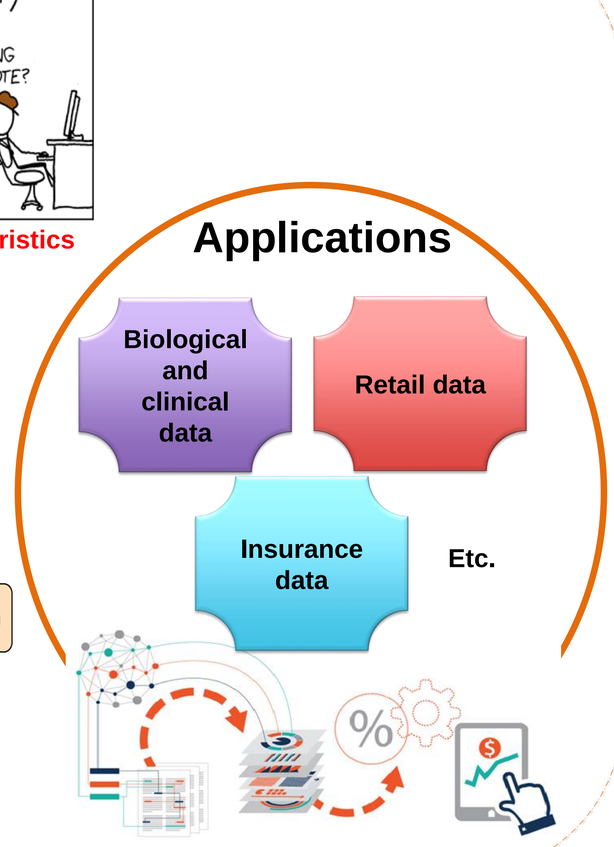
1) The framework provides foundation for future development of improved analysis tools for big data mining and feature selection in certain and imprecise contexts.
2) We develop automated dimensionality reduction techniques, without requiring extra information and with less information loss, which improves the state-of-the-art methods.
3) We develop optimised methodologies to deal with the big data feature selection task with and without the big data veracity aspect, which was not done before. These methodologies use powerful randomised search heuristics for a fast, accurate and a high quality system response.



Apache Spark is a lightning-fast cluster computing technology, designed for fast computation



Randomised search heuristics



Application for marketing

The data set was derived from customer reviews on the Amazon commerce website by identifying a set of most active users and with the goal to perform authorship identification.

To reduce the computational effort of the rough set computations, our approach splits the given dataset into partitions with smaller numbers of features which are then processed in parallel.

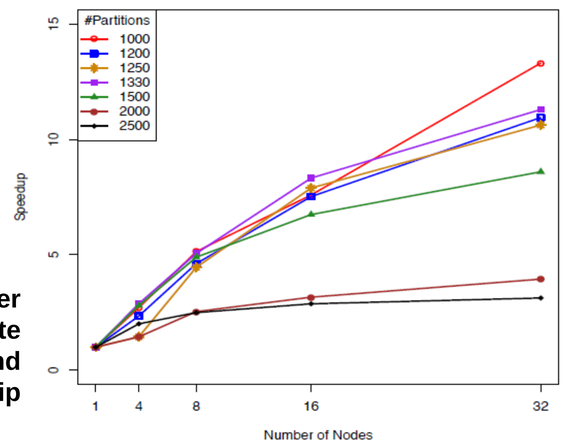


Fig. 1. Speedup for different numbers of nodes and partitions.

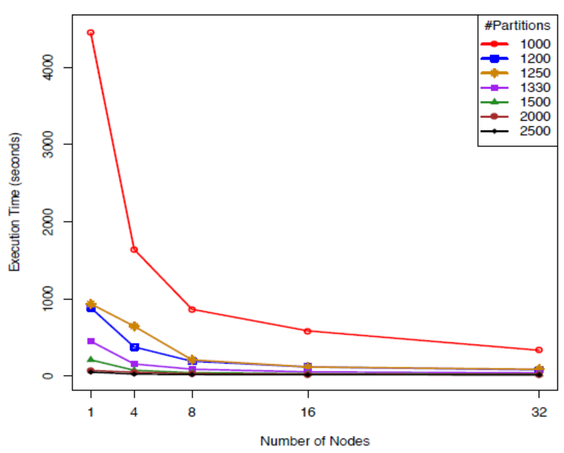


Fig. 2. Average execution times

- ✓ **Speedup and execution time:** There is some trade-off between the number of partitions and the number of nodes used. If only few nodes are available, it may be advisable to use a larger number of partitions to reduce execution times while the number of partitions becomes less important if a high degree of parallelization can be offered.
- ✓ **Stability of Feature Selection:** Our method is very reliable in identifying features for removal.
- ✓ **Classification error with/without our solution:** Our method performs well its feature selection task without any significant information loss.
- ✓ **Execution time with and without our proposed solution:** The overall execution time is decreasing for increasing number of partitions.

Publications:

- *Zaineb Chelly Dagdia, Christine Zarges, Gael Beck and Mustapha Lebbah, "A Distributed Rough Set Theory based Algorithm for an Efficient Big Data Pre-processing under the Spark Framework". IEEE BigData'2017, Boston, USA.*
- *Zaineb Chelly Dagdia, Christine Zarges, Gaël Beck and Mustapha Lebbah, "Modèle de Sélection de Caractéristiques pour les Données Massives", Proceedings of the 15ème édition de l'atelier Fouille de Données Complexes, FDC'2018, Paris, France.*
- *Zaineb Chelly Dagdia, Christine Zarges, Gaël Beck and Mustapha Lebbah, "Nouveau Modèle de Sélection de Caractéristiques basé sur la Théorie des Ensembles Approximatifs pour les Données Massives", Proceedings of the 18ème édition de la conférence internationale francophone Extraction et Gestion de Connaissances, EGC'2018, Paris, France.*

Conclusion:

The project develops a Big Data Mining Framework combining three major fields: "Machine Learning", "Optimization" and "Big Bata"; offering huge opportunities to different communities. It is a unique possibility to explore the impact of such hybridizations within a real-world-application, made possible by collaboration with a pioneering industrial partner.

Current and future work:

- ❑ Analyse further the developed algorithm.
- ❑ Extend the algorithm to deal with clustering application areas.
- ❑ Formalise and implement an optimised version of the framework.
- ❑ Derive a general formulation of rough sets to handle the veracity aspect.
- ❑ Demonstrate the novel methodology on real-world data.

Acknowledgment:

This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 702527.